Research and Methodology Directorate

U.S. Census Bureau Reidentification Studies

By Laura McKenna

Issued April 2019

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

INTRODUCTION[1]

The U.S. Census Bureau conducts its censuses and surveys under Title 13, U.S. Code, Section 9 mandate to not "use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports (13 U.S.C. § 9 (2007))." The Census Bureau applies Disclosure Avoidance (DA) techniques to its publicly released statistical products in order to protect the confidentiality of its respondents and their data. After DA techniques are employed, it can be useful to conduct a motivated intruder reidentification study to assess the disclosure risk of microdata and tabular data products before they are made publicly available. [redacted]

For microdata, such reidentification studies are performed by looking for unique combinations of variables in the microdata that are thought to be identifying, looking for externally available data sets that contain the same variables, and then linking data records in the two data sets using the linkage variables. Finally, it is necessary to verify the proposed matches by comparing the suppressed identities in the microdata with the identities in the external data set to see if the matches are true matches or false matches. This last comparison step is vital, because often survey records are unique within the sample but not in the population (Ramachandran, 2012). A few small reidentification attempts were made with microdata files by summer interns in the early 1990s, but they yielded nothing of substance. The most recent reidentification study for microdata at the Census Bureau was done for the American Housing Survey (AHS) public-use microdata file, which is funded by the Department of Housing and Urban Development (HUD).

For tabular data, reidentification studies often attempt to link tables produced from a given survey or census. The goal is to determine if there are cells appearing in several tables that could be linked together to form microdata records for people or households in small geographic areas. The most recent (completed) reidentification study for tables at the Census Bureau was done for the American Community Survey (ACS) special tabulations to be produced for the Census Transportation Planning Products (CTPP) funded by the American Association of State Highway and Transportation Officials (AASHTO).

The reidentification studies described below are the only Census Bureau studies of which the author is aware.

SURVEY OF INCOME AND PROGRAM PARTICIPATION (SIPP)

In 2000, Dr. Latanya Sweeney (then at Carnegie Mellon) was working on disclosure risk issues for data. She was aware of the Census Bureau's efforts to publish as much high-quality data as possible while maintaining the confidentiality of respondents and their data. She informed the Census Bureau that she had identified two housing units (with two people in each of them) on a SIPP public-use microdata file (PUF). She supplied the names of the four people in the two households to the Census Bureau, and she had correctly identified them. She was able to identify them with the use of public records and media such as newspapers. She said these were just two examples and that she suspected that a vast number of other records could be identified. This greatly alarmed the Census Bureau executive staff, the Disclosure Review Board[2] (DRB), and staff doing research in the field of DA.

Dr. Sweeney then visited the Census Bureau to describe how she identified the two households in SIPP. Both households consisted of elderly married couples in the sample, but while in sample, the make-up of the households changed to one widowed woman. Obviously, the husbands had passed away while the households were in sample. Dr. Sweeney then scanned death records and obituaries in newspapers to identify the two couples. The Census Bureau hired Dr. Sweeney and a couple of her graduate students to reidentify as many records as possible on the same SIPP microdata file, and send the Census Bureau a list of names of those she believed she had identified to staff members who would then identify those that were correct. Her first estimate of the number of people she could identify was 40,000 (again greatly alarming Census Bureau staff).

The DRB immediately issued a confidential addendum to their checklist addressing this problem by adding

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

[2] Census Bureau data products must be approved before dissemination by the DRB.

a small amount of random noise to the ages of elderly married couples, and waiting a few random months before showing births, deaths, marriages, and divorces. The addendum addressed a few other issues including presence of triplets or more, women that gave birth at unusual ages (unusually young or old), households with a very large amount of people, and housing types that were unusual for a given area.

As time passed, the estimate of 40,000 dropped dramatically (to 4,000, then 400, then 40). No additional names were sent after those first four. Dr. Sweeney never fulfilled her contract. She was contacted repeatedly for updates and/or a list of names of reidentified people. The work was not done and she was not paid.

The addendum (latest version in Attachment A) was a benefit that came out of this scare, and another bonus was that the Data Stewardship Executive Policy Committee (DSEP) was formed in 2001. The DSEP ensures the Census Bureau maintains its commitment to protect the confidentiality of respondent's information by fulfilling the legal, ethical, and reporting obligations levied by Title 13 of the U.S. Code. The DSEP is the focal point for decision-making and communication on policy issues related to privacy, security, confidentiality, and administrative records. It oversees several staff committees, such as the DRB, that focus on these important issues. It acts on behalf of the full executive staff in setting policy and making decisions on policy-related matters within the scope of the committee.

CENSUS TRANSPORTATION PLANNING PRODUCTS

The CTPP is a set of ACS special tabulations funded by AASHTO. In 2008, AASHTO presented the DRB with the table shells that they proposed for the next CTPP release. The DRB did not approve this request. One variable, Means of Transportation, often called Mode, was in 18 residence tables (as well as several workplace tables and residence-to-workplace flow tables), and in each of those tables, Mode was crossed by a different variable, e.g., Mode by Age, Mode by Sex, Mode by Occupation. The danger was that the same weighted value might appear for a given mode in all of the tables. If that was the case, it is highly likely that the weight represented one person. A data user could see that the same weight was present for that mode in all of the tables and then link the values of the variables that were crossed with mode to form a microdata record for that person in that geographic area. The Census Bureau does not release any microdata files that identify any areas with a population less than 100,000. The areas for which CTPP tables are published are much smaller than that, for example tracts, places, counties, and traffic analysis zones. Thus, the values linked in the different tables to form microdata records would essentially be published for very small areas. The CTPP request was denied by the DRB. The DRB recommended that AASHTO consider dropping some tables altogether or suppressing the tables where this problem occurred.

AASHTO disagreed with the DRB ruling and met with the Census Bureau executive staff and the DRB to appeal the decision. See a brief write-up prepared for this meeting in Appendix B. Census Bureau DA researchers showed a specific example of how microdata records could easily be formed from the CTPP proposed tables, so everyone involved knew that the CTPP request in its present form could not be approved, but all involved agreed that they would work together to find a solution to this problem.

The National Academy of Sciences (NAS) brought the National Cooperative Highway Research Program (NCHRP), AASHTO, and the Census Bureau together and put out a call for contractors to work with all involved to develop a method(s) to protect the data and still publish most of the desired tables. WESTAT was funded for this work by the NCHRP (Krenzke et al., 2011; 2013; 2017). The WESTAT principle investigator was Tom Krenzke. Census Bureau facilitators were Laura McKenna and Brian McKenzie. The program manager from AASHTO was Penelope Weinberger. The project needed to be completed in 2011 for the 2008 to 2010 3-year ACS CTPP data to be released on schedule.

The project was very successful in that all parties were satisfied with the tables released, the data quality, and the data confidentiality protection, which included a number of different DA techniques (Krenzke et al., 2011; 2013; 2017).

ATTEMPT TO LINK ACS PUBLIC-USE MICRODATA FILES TO OUTSIDE PUBLIC DATA FILES WITH IDENTIFIERS

In 2012, researchers at the Census Bureau, Georgetown, and Harvard Universities made such an attempt. They used what they called identifying attributes on the ACS PUFs to see if they could reveal any other potentially sensitive information about a given respondent represented in a file (Ramachandran et al., 2012). There were two parts to their study. The first was an attempt to link an ACS PUF to public-use data that could be purchased from wholesale data sellers. The second was an attempt to link two profiles on two different networking sites, Facebook and Twitter (not in scope for this paper). They were not very successful linking the ACS PUF to an outside file. They were successful in linking data from the different networking sites, though it took quite a bit of effort and the linkages were specifically targeted, not random. The researchers combined two reidentification strategies and described the results <www.census.gov/srd/papers/pdf/rrs2012-13.pdf>. They then attempted to aid the Census Bureau in identifying variables on the PUF that could lead to reidentification, and discussed a method for doing this using a synthetic data set.

The Reidentification Algorithm is presented in the paper found at the Web site shown above, as are the two reidentification strategies, get combo(s) and check match(es). The ACS study was conducted using data from three U.S. counties and data purchased from wholesale lists that contained identifiers and other variables including ethnicity, gender, age, and income. Finding a data set that could possibly be matched to the ACS proved much more difficult than expected, and most outside data sets had less than 10 variables that overlapped with the ACS PUF. After the study, the researchers concluded that the purchased data was not very accurate. Readers are encouraged to see Table I in the paper for specific results, but in general, the overall vulnerability (correct linkage rate) for the population in the study was less than 0.005 percent. The authors concluded that large-scale reidentification is unlikely when using basic reidentification techniques.

The remainder of the paper discusses linking profiles from Facebook and Twitter, and a method to look for variables that could be causing reidentification problems (explained using a synthetic data set). The researchers did not give a list of such variables for ACS specifically, but at one point in the paper they concluded that for ACS, gender crossed with age lead to most reidentification problems. While gender has just two categories, perhaps age (now in single years) should be collapsed into categories or noise should be added to reduce disclosure risk.

AMERICAN HOUSING SURVEY

In 2013, an external repackager of Census Bureau data brought to the Census Bureau's attention that he reidentified one housing unit record in the New York City Housing Vacancy Survey (NYCHVS) PUF. The Center for Disclosure Avoidance Research (CDAR) confirmed that the suspected reidentification was correct, but the Census Bureau did not inform the repackager of the confirmation. The NYCHVS has many of the same variables as the AHS, but the AHS is a much larger survey. This finding led to a reidentification study on the AHS, using CoreLogic data as a potentially linkable attacker file. Census Bureau staff members (Aref Dajani, Tamara Cole, and their staffs) worked with Dr. Shawn Bucholtz of HUD on this study. The group had to do a lot of work to get to the point where they could try linking an AHS file with CoreLogic data due to differences in definitions and categories of variables.

Once both data sets were ready, linkage attempts were made using three different metrics. The three metrics were called unicity, taxicab (L1 norm), and Euclidean (L2 norm). The unicity metric bins continuous variables and matches records in the attacker and defender file if they have frequency of "1" with respect to any cross-tabulation of variables. The other metrics also bin variables and use the similarities in binned values to create a distance between any attacker and defender record, which is then compared against a cutoff to determine whether any suspected linkages exist. The researchers were interested in three different percentages: the percentage of records that were suspected linkages, the percentage of records that were confirmed linkages, and the conditional rate, which is the percentage of suspected linkages that were confirmed. Researchers set a very low threshold for the conditional rate they felt was acceptable.

In the 2013 investigations, the suspected rates of reidentification using a unicity attack varied widely, and the confirmed and conditional rates were zero. However, for the taxicab and Euclidean attacks, all three rates varied widely. For these two metrics, the researchers often uncovered thousands of attacker units linked to every defending unit with a confirmed reidentification, leading to a focus on defending units that matched to five or fewer attacking units. In a study of the 2015 AHS, the taxicab metric resulted in conditional reidentification rates of zero for three of 12 metro areas examined, less than the threshold for one additional area, slightly above the threshold for six areas, and far above the threshold for two areas. The Euclidean rates showed similar results. Based on the conditional reidentification rate of 2014 CoreLogic data and a noise-infused and collapsed 2015 AHS PUF, CDAR recommended that the PUF not be released without further protection. As a result, a few variables were dropped from the PUF, and researchers are now testing noise-infusion techniques that can be applied to those variables so that they may be again included on future AHS PUFs.

CONCLUSION

Reidentification studies have proven helpful to the Census Bureau in the past. No matter what the reidentification rates were, the studies' results have shown the Census Bureau where changes should be made to public files, where disclosure risks are present, and which files need additional protection.

John Abowd, Associate Director for Research and Methodology (Census Bureau), is currently leading a reidentification study similar to the CTPP study, but on a much larger scale. It is similar to the CTPP study in that the disclosure risk lies with the ability of users to link tables together in order to form microdata records for very small areas from the decennial census. There are currently no additional reidentification studies planned.

Recently, the Census Bureau has embarked on an aggressive effort to replace its legacy DA methods with modern DA techniques based on formal privacy methods <https://privacytools.seas.harvard.edu/formal-privacy-models-and-title-13>. Current methods will gradually change with the introduction of formal privacy (Nissim et al., 2018). Most of the current Census Bureau's DA research is focused on formal privacy for all types of data (Nissim et al., 2007). An algorithm operating on a private database of records satisfies formal privacy if its outputs are insensitive to the presence or absence of any single record in the input (Dwork, 2006). The DRB is quickly learning about formal privacy and how it protects Census Bureau data products. Because of this, the Census Bureau may or may not need to conduct additional reidentification studies.

REFERENCES

C. Dwork, "Differential Privacy," International Colloquium on Automata, Languages, and Programming (ICALP), 2006, pp. 1–12.

T. Krenzke, J. Li, and L. McKenna, "Producing multiple tables for small areas with confidentiality protection," *Journal of the International Association of Official Statistics*, Volume 33, No. 2, 2017, pp. 469–485. Doi:10.3233/SJI-160259.

T. Krenzke, J. Li, and L. Zayayz, "Balancing use of weights, predictions and locality effects in a model-assisted constrained hot deck approach for random perturbation," Proceedings of the Survey in Research Methods Section of the American Statistical Association, Alexandria, VA, 2013.

T. Krenzke, J. Li, M. Freedman, D. Judkins, D. Hubble, R. Roisman, and M. Larsen, "Producing Transportation Data Products from the American Community Survey that comply with disclosure rules," National Cooperative Highway Research Program, Transportation Research Board, National Academy of Sciences, Washington, DC, 2011.

K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth Sensitivity and Sampling in Private Data Analysis," Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing, 2007, pp. 75–84.

K. Nissim, T. Steinke, A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, D. O'Brien, and S. Vadhan, "Differential Privacy: A Primer for a Non-technical Audience (Preliminary Version)," Harvard University Privacy Tools for Sharing Research Data, 2018, <http://privacytools.seas.harvard.edu>.

A. Ramachandran, L. Singh, E. Porter, and F. Nagle, "Exploring Re-Identification Risks in Public Domains," Tenth Annual International Conference on Privacy, Security and Trust, Institute of Electrical and Electronics Engineers, Danvers, MA, 2012, DOI:10.1109/pst.2012.6297917.

Appendix A: [redacted] Addendum to the U.S. Census Bureau Disclosure Review Board Checklist

Revised October 6, 2009

ADDITIONAL DISCLOSURE METHODS FOR PUBLIC-USE MICRODATA FROM DEMOGRAPHIC SURVEYS

In addition to the population requirement of 100,000 people in each identified geographic area (or higher, as dictated by content and available information on sample design) and the application of 3%/0.5% topcoding rule, demographic surveys need to employ some other DA procedures. First, we must ensure observable demographic events and housing characteristics are protected from disclosure. Second, potential events (such as deaths) that could contribute to the reidentification of a respondent must be considered in protecting the data for release. Third, public-use files need to be processed through the most recent reidentification software to ensure no disclosures are feasible.

Observable Demographic Events

Event 1: Difference in age between mother and natural-born children

For children born in [redacted] when there are two or more natural-born children where the difference in ages between the mother and children is less than [redacted] or when there are [redacted] children where the difference in the age between the mother and child is greater than [redacted] the new disclosure method should be selected from one of the following:

- Relocate the housing unit to another state.
- Perturb year of birth (or age) for both mother and child(ren).
- Recode relationship codes in these vulnerable situations so that the parent/natural-born child relationship is not revealed.

It is understood that a small number of ages cannot be perturbed. For those that can, perturbations of zero should not be allowed. When ages are perturbed, an attempt should be made to conform to the particular survey or Census Bureau edit procedures.

Event 2: Simultaneous births

Regardless of presence of parent(s) and regardless of age, for [redacted] simultaneous births [redacted] [redacted], the new disclosure method should be selected from one of the following:

1. Remove housing units with [redacted] simultaneous births, and randomly relocate housing units with [redacted] to another state.
2. Perturb year of birth (or age) so that the apparent number of simultaneous births is reduced [redacted] [redacted]
3. ONLY to be used if both Options #1 and #2 are not viable for a particular set of n-tuplets, then select from one of the following to reduce the number of apparent simultaneous births:
   a. Recode relationship codes in these vulnerable situations [redacted]
   b. Perturb year of birth (or age) of [redacted] [redacted].

Option 3a is preferable over Option 3b because of the increased protection given to the n-tuplets. However, Option 1 or Option 2 are the most preferred options and should be used whenever possible.

It is understood that a small number of ages cannot be perturbed. For those that can, perturbations of zero should not be allowed. When ages are perturbed, an attempt should be made to conform to the particular survey or Census Bureau edit procedures.

Event 3: Births, deaths, marriages, and divorces

For demographic events, such as births, deaths, marriages, and divorces, the [redacted] that the event takes place should be perturbed (this is mainly for longitudinal surveys).

Observable Unusual Housing Characteristics

For persons residing in unique structures as determined by characteristics of the dwelling and location (for example a mobile home in Washington, DC), select one the following:

- Mask or suppress information on visible characteristics of the dwelling (e.g., number of units in the structure, number of floors, type of dwelling unit).
- Delete geographic detail within which visual characteristics become unique, e.g., if there is only one high-rise apartment building in a particular area, either do not identify it as a high-rise, do not identify that area, or move the housing unit to another state.

Potential Demographic Events

For longitudinal surveys, extra care must be taken to protect the confidentiality of information concerning people who die while in the survey. Program areas need to mask the death of a person. [redacted]

Attacks

The Census Bureau will be continuously monitoring techniques and data that can be used to break Census Bureau disclosure methods. This research will result in software that can measure the vulnerability of a given file before release. When such software is available, all public-use files will be checked through this software at the point where DRB review is warranted. When new techniques or data are added to the archive maintained by the unit that does this monitoring, surveys may be required to undergo DRB review, even if nothing else has changed that would warrant such review.

Appendix B: The U.S. Census Bureau Disclosure Review Board's Decision on the Census Transportation Planning Products Proposal for Special Tabulations of the American Community Survey's 3-Year Estimates

PREVIOUS RULINGS

The Disclosure Review Board (DRB) has made a small number of previous rulings on standard (often called "base") transportation-related tables and Census Transportation Planning Products (CTPP) special tabulations. Previous rulings were for the 2000 Census base and CTPP special tabulations, for the American Community Survey (ACS) base tables, and for 1-year data ACS CTPP special tabulations, but not for 3-year CTPP special tabulations that would be released in addition to the base tables. The geographic areas in the 1-year data tables all represent areas with a population of 65,000 people or more. None of those previous requests were the same as the current request, which is for CTPP special tabulations for 3-year data for areas with a population of 20,000 people or more.

THE "THRESHOLD OF THREE" REQUIREMENT

The DRB ruling on the CTPP 3-year data request required a threshold of three unweighted cases for each given category of means of transportation in each geographic area for tables where that variable is crossed with one or more other variables. Note that there is no threshold for the univariate table showing means of transportation (crossed with no other variables). The use of a threshold of three unweighted cases for universes, variable categories, and even individual tables cells is quite common for publicly released data products. The DRB has used the threshold of three on variables such as respondents speaking a given language, foreign-born noncitizens, migration flows, mail eligible housing units, certain ancestries, establishments (for economic data), etc. The DRB set this criterion for CTPP because of the large number of two- and three-dimensional tables that include means of transportation as one of the variables (18 residence tables, 17 workplace tables, and 6 flow tables). If a person is the only person in sample with a given means of transportation in their area of residence, area of workplace, or flow, tables can be linked together to form a microdata record for that person who lives in that area with a population of 20,000 people or more. We do not want this to happen, thus the threshold.

THE "THRESHOLD OF THREE" EXAMPLE

For example, there may be a person who is the only person in sample who rides a bicycle to work in some geographic area. A univariate table shows a weighted count of the number of people who rode a bicycle to work in that (say) county, and let's say his weight is 30. A data user would see a 30 in the table cell of people who rode a bicycle to work in that county. There is also a table that includes the weighted number of people who rode a bicycle to work by occupation. There would be one occupation category for people who rode a bicycle that shows a weighted value of 30 and the rest of the occupation categories are zeros. There is also a table that includes the number of people who rode a bicycle to work by income. There is one income category that shows a weighted value of 30 for people who rode a bicycle and the rest of the income categories are zeros. There is also a table that includes the number of people who rode a bicycle to work by race. There is one race category that shows a weighted value of 30 for people who rode a bicycle to work and the rest of the race categories are zero. There are 18 tables like that in the request for tables with areas of residence (the same thing can happen for workplace and flow tables). The person with the weight of 30 is clearly one person. His 18 characteristics can easily be linked and form a microdata record for a small geographic area (population 20,000 or greater). Note that this record represents all of his data at one point in time (a snapshot of when he took the survey). That is a microdata record for an area much too small. We have asked Douglas Hillmer, from the American Community Survey Office, for data to show examples of how data items from tables can be linked together to form microdata records. (For the meeting, Paul Massell put together and showed a true example of this).

GEOGRAPHIC REQUIREMENTS FOR MICRODATA

We cannot release microdata records (or tabular data that can form microdata records) for areas with a population of only 20,000 people. In 2002, the Disclosure Avoidance Research Group in the Statistical Research Division started doing reidentification studies where we compare non-Census Bureau publicly available data products with Census Bureau publicly released data products to ensure there are no disclosure problems. We already have to find and fix problems in microdata files that identify areas with populations of 100,000 and 250,000. Statistical Policy Working Paper 22 published by the Federal Committee on Statistical Methodology shows that no U.S. federal statistical agencies publish microdata for areas with less than a population of 100,000 and most have a much higher threshold (state, for example). This criterion is the same for statistical agencies around the world.

BASE TABLES, DATA QUALITY, AND OPTIONS FOR COLLAPSING CATEGORIES

The DRB realizes the importance of data on means of transportation. This variable is already crossed with 14 other variables in the ACS base residence tables and the same 14 for the base workplace tables. The DRB only agreed to this because those tables must pass the ACS data quality filter, which also helps with DA.

Taken from "Design and Methodology" Technical Paper 67 available at <www.census.gov/acs/www/Downloads/tp67.pdf>:

> "Even with the population size thresholds described earlier, in certain geographic areas some very detailed tables might include estimates whose reliability is unacceptable. Data release rules, based on the statistical reliability of the survey estimates will be used starting with the 2005 ACS data released in the summer of 2006. These release rules apply only to the single-year and 3-year data products.
>
> "The main data release rule for the ACS tables works as follows. Every base table consists of a series of estimates. If more than half the estimates are not statistically different from zero (at a 90 percent confidence level), then the table fails. Each estimate is subject to sampling variability that can be summarized by its standard error. Dividing the standard error by the estimate yields the coefficient of variation (CV) for each of the estimates. (If the estimate is zero, a CV of 100 percent is assigned.) To implement this requirement for each table at a given geographic area, CVs are calculated for each of the table's estimates, and the median CV value is determined. If the median CV value for the table is less than or equal to 61 percent, the table passes for that geographic area; if it is greater than 61 percent, the table fails. Tables that are too sparse will fail this test. In that case, the table will not be published for the geographic area.
>
> "Whenever a table fails, a simpler table that collapses some of the detailed lines together can be substituted for the original, more detailed table. The data release rules are then applied to the simpler table. If it passes, the simpler table is released. If it fails, none of the estimates for that particular table is released for this geographic area. These release rules are applied to single-year period estimates and multiyear estimates based on 3 years of sample data. No data release rules are applied to the estimates based on 5 years of sample data."

The DRB feels that crossing means of transportation with 18 variables (using residence tables as the example) without the quality filter would pose an unacceptable disclosure risk. Those requesting the data note that the 82 percent table suppression (due to the DRB threshold) can be reduced to 33 percent table suppression if they use a collapsed version of means of transportation. The DRB has informed them that they may collapse categories of means of transportation or collapse geographic areas any way they wish to in order to meet the threshold. Letting data users form microdata records from 82 percent of the requested tables is unacceptable.

A SNAPSHOT OF A PERSON AT ONE POINT IN TIME

Those requesting the data note that we use data swapping to protect data, some of the data may be a few years old, and some information for a given respondent may have changed. Data swapping targets unique records, but means of transportation is not one of the key variables used to find unique records (we use other more publicly available variables). So having a unique means of transportation does not mean that someone's record will be swapped, but a unique means of transportation can certainly be used to link other data variables together from tables to form a microdata record. While some variables can change over time (marital status, occupation, age group, etc.), when a set of tables shows that one person in sample in one area had a given means of transportation and you can form their microdata record, that is a snapshot of that one person at a given time. That is a microdata record for that person at a given time that was not that long ago (say 3 years) in an area with a population of 20,000. As noted earlier, the smallest population of areas for which microdata records are released is 100,000, and sometimes the threshold is 250,000 or higher.

DRB EFFORT TO MAKE DATA AVAILABLE

Those requesting the CTPP special tabulation took minutes from previous meetings with DRB members. Those minutes show that we repeatedly told them that if they crossed means of transportation with many other variables, there would be a threshold of three unweighted cases applied to that variable's categories in each geographic area. After the DRB discussed their proposal and drafted minutes and a memorandum on the ruling, the DRB gave them 1 month to respond and perhaps revise their request before issuing the formal memo. They did not respond until the memo was issued. The DRB gave those requesting the data an option of using a synthetic data technique that Nanda Srinivasan (now at NAS) developed for them, but they said they do not have enough data to use to develop a synthetic 3-year product. They will need to wait for the 5-year product. We asked Nanda Srinivasan about the possibility of altering weights to prevent users from forming microdata records, but there probably is no time to determine if this is a feasible alternative. Nanda Srinivasan acknowledges that microdata records can be formed from the tables.

CENSUS BUREAU EFFORT TO MAKE DATA AVAILABLE:

Finally, it is debatable as to whether those requesting base data tabulations on transportation, as well as the CTPP 3-year data product, are getting more or less information than was previously made available. From the 1990 Census, not much was provided (base tables) or requested (special tabulations). Many, but not all, of the tables that they are requesting in this ACS 3-year special tabulation were provided in a 2000 Census special tabulation. This was a decision made before the Census Bureau began reidentification studies, and it was for a request for tables that included 17 percent versus 7.5 percent of the population. A larger sample yields a smaller percentage of sample uniques (Zayatz, 1991). Also, the 2000 Census base tables only crossed means of transportation with two variables: travel time to work and race. As stated previously, the ACS base tables cross means of transportation with 14 variables, so the users can get many more base tables (with no charge) on a routine basis as long as those tables pass the data quality filter.

REFERENCE

L. V. Zayatz, "Estimation of the Percent of Unique Population Elements on a Microdata File Using the Sample," RR-91-08, *Statistical Research Division Report Series*, U.S. Census Bureau, 1991.